Loop Problem in Proteins: Developments on the Monte Carlo Simulated Annealing Approach

LOUIS CARLACCI*

Department of Chemistry, University of South Florida, Tampa, Florida 33620-5250

S. WALTER ENGLANDER

The Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, School of Medicine, Philadelphia, Pennsylvania 19104-6059

Received 4 October 1994; accepted 16 August 1995

ABSTRACT

Calculations of loop segments in bovine pancreatic trypsin inhibitor starting from random conformations are more efficient, reproducible, and reliable due to several program enhancements. Monte Carlo simulated annealing (MCSA) calculations of a five-residue α-helix N-terminus segment (H5) and β-strand segment (B5) in this study are compared to the corresponding loop calculations in our previous study. Characteristics of the calculations are: the lowest final total energy conformations (LECs) are within 5 kcal/mol; the average backbone deviations of the computed segments from the native X-ray conformations are 0.43 ± 0.15 Å for H5 and 0.68 ± 0.20 Å for B5; and all the native backbone–backbone hydrogen bonds (H bonds) are present in the best LECs. Compared to the previous study, the H5 and B5 calculations are about 3 and 24 times more efficient, respectively. In the analysis of the best H5 simulated annealing run, backbone–backbone H bonds appear between RT = 4 and 70 kcal/mol, where RT is the Boltzmann temperature factor. H bonds that involve side chains appear in the RT = 1–10 kcal/mol range. Program enhancements implemented are varying main chain versus side chain dihedral angle selection rate, varying ϕ/ψ and χ_1/χ_2 dihedral angles in pairs, the use of main chain and side chain rotamer libraries, and varying the location of the segment origin.

* Author to whom all correspondence should be addressed.

CCC 0192-8651 / 96 / 081002-11

Introduction

The multiple minimum problem in protein conformational search calculations arises from the many degrees of freedom and the complex potential energy function employed. The problem is especially apparent in flexible loops on the protein surface, for example, the loops that connect secondary structure elements in protein tertiary fold motifs (4-α-helix bundle,[1-3] $\beta\alpha\beta\alpha\beta$ crossover,[3-5] β barrels[3,6-10]), the loops that connect the membrane spanning regions in membrane protein,[11] and the transient nonhydrogen bonded loops that lead to hydrogen exchange with solvent.[12,13] For loops stabilized by hydrogen bonds (H bonds), hydrophobic interactions, electrostatic interactions, and/or disulfide bridges, the Boltzmann distribution of loop conformations consists of a limited number of distinct low energy conformations. For flexible surface loops, the Boltzmann distribution consists of many distinct low energy loop conformations. In this study, surface loops with many H bonds are computed starting from random conformations.

There are several implementations of the MCSA method in protein structure calculations.[14-20] In an MCSA simulation, the folding pathway followed can take high energy as well as low energy routes. The random walk method used to sample conformations allows the conformational search to tunnel through energy barriers. In an MCSA cycle, a trial conformation is evaluated based on a statistical comparison with the currently accepted conformation. The Boltzmann distribution temperature factor (RT, in kcal/mol) is lowered stepwise during the course of simulated annealing. The affect of lowering RT is to confine the search for the global energy minimum to lower energy folding pathways.

Higo et al.[18] and Collura et al.[19] employed the MCSA approach for the calculation of loops starting from the fully extended conformation. The MCSA calculation by Carlacci and Englander[20] started from random loop conformations. In all three studies, the MCSA calculation considered just the loop while the surrounding local region that interacts with the loop remained fixed. The surrounding local region can be considered a template on which the detached loop can dock. Higo et al.[18] computed three systems of complementary determining regions of antibodies. Root mean square (rms) backbone deviations between heavy atoms of the computed loops and heavy atoms of the native X-ray conformation (referred to as rms backbone deviations) are 1.81 Å for a seven-residue loop, 2.47 Å for the simultaneous calculation of five- and nine-residue loops, and 2.11 Å for a five- and seven-residue loop pair. Collura et al.[19] computed loop segments in the immunoglobulin McPC603, bovine pancreatic trypsin inhibitor (BPTI), and bovine trypsin. The average rms backbone deviation of the computed seven-residue N-terminus helical segment in BPTI, which overlaps the helical segment computed in this study, is 0.5 Å. In our previous study[20] the rms backbone deviations were in the range 0.9–1.1 Å for five-residue segments (α-helix N-terminus segment, β-strand segment, and loop segment), and 1.6 and 1.9 Å for seven- and nine-residue loops. Loop calculations by Bruccolieri and Karplus[21] used various filters in a procedure that uniformly samples conformational space. Fine and collaborators[22,23] employed energy minimization and molecular dynamics to predict the conformations of surface loops starting from many initial conformations. Energy minimization starting from many arbitrary loop conformations was used to compute surface loops that connect secondary structures in various types of protein tertiary fold motifs.[24-29] The difficulty in all the above approaches to the calculation of loop conformations is that the computational effort required increases nonlinearly with loop size and with distance between the loop ends.

In this study, loop conformations are predicted using the MCSA approach starting from random conformations. Only the part of the protein that interacts with the loop is considered in energy calculations. Equilibration runs which force the simulation to spend more time at high RT are used to remove bad contacts in the random initial loop conformation. Each equilibration run is followed by multiple independent production runs of limited extent. An enhanced simulated annealing program (called Safd/2), which is an updated version of program Safd,[20] is used to calculate loops in BPTI. The program enhancements implemented are varying the rate that main chain and side chain dihedral angles are varied, the selection of main chain ϕ/ψ and side chain χ_1/χ_2 dihedral angles in pairs, the use of a side chain rotamer library,[30,31] the use of a main chain ϕ/ψ rotamer library,[32] and varying the location of the segment origin throughout the simulation.

Our goals in this study are to show that the loop calculations are more efficient and accurate compared to our previous study,[20] and to determine the properties of the best simulated annealing runs. The best simulation run of the α-helix N-terminus segment in BPTI is characterized in terms of H-bond formation, molecular energetics, and acceptance ratio during the simulated annealing.

Methodology

The native conformation of a BPTI loop is defined here as the conformation of the loop in the X-ray crystal structure called 4pti in the Brookhaven Protein Data Bank.[33] In the analysis of the computed loop, the part of the protein that remains fixed during the simulation is overlayed on the native X-ray structure; and the rms deviation of the computed loop from the native is determined. The computed loop with the lowest final total energy (called the LEC) is considered the best conformation. The IUPAC nomenclature for amino acids[34] is employed.

LOCAL REGION

The local region consists of the computed segment and selected nearby residues that interact with the computed segment (called the surrounding local region). At the end of a trial and error selection process, the surrounding local region consists of residues within 8 Å of the computed segment.

The computed β-strand and α-helix N-terminus segments in BPTI are named B5 and H5, respectively. The B5 segment is a five-residue segment that begins at residue 16 and ends at residue 20. The amino acid sequence is Ala-Arg-Ile-Ile-Arg. The four surrounding local region segments are residues 9–15, 21–22, 32–40, and 44–46. Five backbone–backbone H bonds (NH to CO) are located on one side of the β-strand segment. The other side of the β-strand segment is exposed on the protein surface. The accessible conformations of the β-strand segment are limited due to the large end-to-end distance of the segment. The H5 segment is a five-residue segment that begins at residue 46 and ends at residue 50. The amino acid sequence is Lys-Ser-Ala-Glu-Asp. The side chain of Arg 53 located on the protein surface is computed as well. The five surrounding local region segments are residues 4–5, 18–23, 29–35, 40–45, and 51–55. H5 has four backbone–backbone H bonds. The short end-to-end distance of the H5 segment relative to the B5 segment makes H5 difficult to compute.

COORDINATES OF LOCAL REGION

Conformations of the computed segment and computed side chains in the surrounding local region are generated from dihedral angles first. Rigid body coordinates, three Euler angles, and three translational coordinates specify the orientation and position relative to the fixed part of the protein. ECEPP[35,36] geometric parameters for L-amino acids are employed. Except for the segment N- and C-termini peptide bonds, which are initially severed from the rest of the protein, all bond lengths and angles are fixed at standard values.

Coordinates of side chains that vary are generated from dihedral angles[35] and rigid body coordinates.[20] The rigid body coordinates are obtained from the transformation that positions atom C^{α} at the origin and that orients atom C^{β} out along the positive x direction and atom N in the xy plane out along the positive y direction.

In this study, the location of the computed segment origin is initialized on a new residue (called the new origin residue) many times throughout the simulation. This is referred to as the origin location option. In the origin location option, Cartesian coordinates of the computed segment are generated as an independent polypeptide chain[35,36] in which the first residue is at the origin. A coordinate transformation locates the segment origin on the new origin residue. Rigid body coordinates, which are parameters in the MCSA protocol, are then used to orient and position the segment in the protein. Each time a new origin residue is selected, every 200 trial conformations, rigid body coordinates that orient and position the segment in the protein are extracted from the current Cartesian coordinates of the segment in the protein.[20] When the first residue is the new origin residue, a main chain perturbation at the N-terminus greatly affects the conformation of the segment C-terminus. When a middle residue is the new origin residue, a main chain perturbation on the N-terminus affects the conformation of atoms only on the N-terminus side; and rigid body coordinate perturbations produce smaller displacements of the segment C-terminus.

EMPIRICAL PARAMETERS AND TOTAL ENERGY

The conformational energy is the sum of electrostatic, nonbonded, and torsional energies. CHARMm (all atom) energy parameters[37] are employed. A hydrogen bond energy is built into the electrostatic term which is given by a Coulomb potential. A distant dependent dielectric is used to screen charges of ionized states of side chains. The calculation neglects explicit water.

The total energy is the sum of the conformational energy of the local region and a weighted polypeptide chain continuity constraint.[20] The total energy is

$$E_{\text{total}} = E_{\text{local}} + \text{XLOOP} \times F_{\text{overlap}} \quad (1)$$

where E_{local} is the conformational energy of the local region and XLOOP $\times$ F_{overlap} is the weighted polypeptide chain continuity constraint. The energy of the local region is

$$E_{\text{local}} = E_{\text{intra}}(\text{CS}) + E_{\text{intra}}(\text{SURR}) + E_{\text{inter}}(\text{CS/CS}) + E_{\text{inter}}(\text{CS/SURR}) + E_{\text{inter}}(\text{SURR/SURR}) \quad (2)$$

where E_{intra} is the sum of all intrasegment energies, E_{inter} is the sum of all interaction energies between two groups of segments, CS represents computed segments, and SURR represents surrounding local region segments. The polypeptide chain continuity constraint is based on the overlap of dummy residues appended to the computed segment with identical residues (called overlapping residues) immediately adjacent to the computed segment in the surrounding local region.[24] The atoms overlapped are backbone atoms N, C^{α}, C′, and O. The polypeptide chain continuity constraint is given by

$$F_{\text{overlap}} = \sum_i ([\mathbf{r}(N) - \mathbf{r}^o(N)]^2 + [\mathbf{r}(C^{\alpha}) - \mathbf{r}^o(C^{\alpha})]^2 + [\mathbf{r}(C') - \mathbf{r}^o(C')]^2 + [\mathbf{r}(O) - \mathbf{r}^o(O)]^2) \quad (3)$$

where i is a dummy residue index, $\mathbf{r}(X)$ is the position vector of atom X in the dummy residue, and $\mathbf{r}^o(X)$ is the position vector of atom X in the overlapping residue. The loop closing penalty, XLOOP, is 50 kcal mol^{-1} Å^{-2}.

The rms overlap deviation[20] is a measure of the smoothness of the connection between the computed segment and immediately adjacent segments. The rms overlap deviation is

$$\text{rms}_{\text{overlap}} = \left[\frac{F_{\text{overlap}}}{4 \times (\text{number of dummy residues})} \right]^{1/2} \quad (4)$$

where four atoms per dummy residue are overlapped. Final conformations with rms overlap deviations greater than 0.1 Å are not characterized in the results section.

MCSA PROTOCOL

The native conformation of BPTI refined by constrained energy minimization (called the refined native conformation) is the starting point of the MCSA calculations. In energy refinement, side chain dihedral angles are optimized first. Next, the main chain is weakly constrained to the native X-ray conformation; and all dihedral angles are simultaneously optimized. MCSA calculations starting from random loop conformations are influenced by the conformation of the local region that the loop rests against.

Several implementations of the MCSA method are found in the literature.[14-20,38-41] In the MCSA method, many trial conformations are sequentially evaluated. At the start of an MCSA cycle, a trial conformation is generated and the total energy is computed. The energy of the trial conformation is compared to the energy of the previously accepted conformation. The trial conformation is accepted if the total energy of the trial conformation is less than or equal to the total energy of the prior conformation. Otherwise, the Metropolis criteria[38] is used to accept or reject the trial conformation. The most recently accepted trial conformation is used in the next cycle. The simulation starts out with a large RT (in kcal/mol). In a simulated annealing step, RT_i, the Boltzmann temperature factor at step i in the annealing, is fixed; and a number of trial conformations (NCON) are evaluated. At the end of each step, the current temperature factor is divided by a factor called β. β is given by

$$\beta = \left\{ \frac{RT_{\text{init}}}{RT_{\text{NSTEP}}} \right\}^{1/(\text{NSTEP}-1)} \quad (5)$$

where RT_{init} is the initial temperature factor, RT_{NSTEP} is the temperature factor of the last step,

and NSTEP − 1 is the number of times that the temperature factor is lowered.

An MCSA calculation consists of three simulation runs, equilibration, prediction, and refinement. A combined prediction plus refinement run is called a production run. Equilibration runs start from random loop conformations. Multiple production runs follow each equilibration run. In the equilibration runs, the initial and final temperature factors, RT_{init} and RT_{10}, are 1000 and 200 kcal/mol, respectively, and NCON is 12,000. In the prediction runs, initial temperature factors are 500, 200, 100, and 75 kcal/mol; and the final temperature factor, RT_{30}, is 1.0 kcal/mol. In the refinement run, the initial and final temperature factors, RT_{init} and RT_{14}, are 1.0 and 0.2 kcal/mol, respectively, and RT_{15} is 0 kcal/mol. NCON of the production run is chosen from experience.

Two seeds are used to generate random numbers. One seed is used to generate trial conformations. The other seed is used to evaluate trial conformations.

Computed segment coordinates that vary are main chain dihedral angles (ϕ, ψ, and ω), side chain dihedral angles, and rigid body coordinates. Select side chain dihedral angles of surrounding local region residues vary. In the initial conformation, side chain dihedral angles with rotamer conformations (described in the next section) are assigned from the rotamer library. ω dihedral angles are initialized to 180°. ϕ, ψ, and χ dihedral angles that remain are assigned random values in the range −180°–180°. Rigid body coordinates are not randomized.

In the random walk method used to generate a trial conformation, a randomly chosen parameter, ang, is assigned a value in the range ang − Pert ≤ ang ≤ ang + Pert, where Pert is the perturbation. Perturbations are defined for main chain and side chain dihedral angles and for Euler angle and translational coordinates. Spherical coordinates, r, ϕ, and θ, define the translation vector. In a translation random walk, ϕ and θ are assigned random values in the full possible range; and r is assigned a value in the range $0 < r \leq$ Pert Å. In equilibration and prediction runs, initial main chain and side chain dihedral angle perturbations are 45° and 180°, respectively; and Euler angle and translation perturbations are 30° and 2 Å, respectively. In refinement runs, initial main chain and side chain dihedral angle perturbations are 5° and 180°, respectively; and initial Euler angle and translation perturbations are 3° and 0.4 Å, respectively.

Trial conformations are accepted or rejected based on the Metropolis criteria.[38] Let ΔE be the difference between the total energy of the trial conformation and the current accepted conformation. If ΔE is less than zero, then the trial conformation is accepted. Otherwise, a random number between zero and one is compared to $\exp^{-\Delta E/RT}$, the Boltzmann distribution. If the random number is less than the Boltzmann distribution, the trial conformation is accepted. Otherwise, the trial conformation is rejected. For a given RT, the probability that a trial conformation is accepted decreases as ΔE increases.

The acceptance ratio is the number of accepted trial conformations divided by the total number of trial conformations evaluated. When the acceptance ratio goes below the acceptance ratio limit, the perturbation used to generate the trial conformation is lowered. Acceptance ratios and acceptance ratio limits (called κ and K, respectively) are defined for various types of trial conformations, namely, trial conformations generated by backbone, side chain, Euler angle, and translation perturbations. Based on experience, K_{backbone}, $K_{\text{side chain}}$, $K_{\text{Euler angle}}$, and $K_{\text{translation}}$ are 25, 40, 25, and 25%, respectively. The factors used to lower the backbone, side chain, Euler angle, and translation perturbations are 1.5, 2, 1.5, and 1.5, respectively. For example, when κ_{backbone} is less than K_{backbone} M times during the simulation, the current backbone perturbation is $\text{Pert}_{\text{backbone}} \div 1.5^M$.

MCSA PROGRAM DEVELOPMENTS

MCSA developments are altering the main chain/side chain dihedral angle selection rate, changing dihedral angles in pairs, and the use of main chain and side chain rotamers. The use of rotamer libraries restricts the conformational search to energetically favorable conformations. (Main chain and side chain rotamer libraries are available on request to the first author.)

The main chain/side chain selection rate option was initiated on the basis of the observation that the main chain acceptance ratio is much lower than the side chain acceptance ratio throughout the simulation. When a dihedral angle is selected, the probability that a main chain trial conformation is generated is increased by 0.15, from 0.5 to 0.65; and the probability that a side chain trial conformation is generated is decreased by 0.15, from 0.5 to 0.35.

In the dihedral angle pairs option, a residue ϕ and ψ (or χ_1 and χ_2) are assigned new values

when either dihedral angle is randomly selected. Varying dihedral angles in pairs allows perturbations in opposite directions, which compensate each other. Certain residues have special considerations. χ_1 and χ_2 do not exist in Gly. Ala does not have a χ_2 dihedral angle. In the Pro residue, ϕ is fixed at 75°; and the side chain does not vary. χ_2 of Ile affects the conformation of the larger branch of the side chain.

The side chain rotamer library is derived from protein X-ray data[30] and small molecule conformational analysis.[31] A rotamer conformation obtained from X-ray data[30] spans the dihedral angle range observed. Rotamer conformations derived from small molecule conformational analysis[31] are: $-CH_3$ and $-NH_3^+$ (tetrahedral: 60° ± 10°); primary amine $-NH_2$, alcohol —OH, and thiol —SH (tetrahedral: −60° ± 10°, 180° ± 10°, and 60° ± 10°); imine $-NH_2$ (planar: 0° ± 5°); carboxylic and phenolic —OH; and secondary amine —NH (planar: 180° ± 5° and 0° ± 5°). In the use of the side chain rotamer library, one of the rotamer conformations is randomly chosen; and a random walk inside the rotamer range is taken. In equilibration and prediction runs, side chain dihedral angles with one rotamer conformation are fixed. In refinement runs, all side chain dihedral angles are varied; and the random walk method is employed when $RT \leq 0.3$ kcal/mol.

Main chain rotamer conformations derived from X-ray data are defined for the dihedral angle ω, and the ϕ/ψ pair. In this study, only the trans rotamer conformation of ω (180° ± 15°) is explored. The ω rotamer option is turned off when $RT \leq 1.0$ kcal/mol. There are 3 ϕ/ψ rotamer classes (called α-helix, β-strand, and β-hairpin). ϕ/ψ rotamer conformations depicted in Figure 1 are based on figure 2 of ref. 32. In Figure 1, the α-helix and β-strand rotamer regions are labeled helix and sheet, respectively; and the five β-hairpin rotamer regions are numbered 1–5. In this study, when either ϕ or ψ is selected, a random walk inside one of the ϕ/ψ rotamer regions is taken 10% of the time; and an unrestricted random walk is followed 90% of the time. In the use of the main chain rotamer library, one of the three rotamer classes is randomly selected first. If the β-hairpin class is chosen, one of the five β-hairpin rotamer conformations is randomly selected next. The random walk perturbation is the smaller of the ϕ/ψ rotamer range and the current main chain perturbation. The ϕ/ψ rotamer option is turned off when $RT \leq 50$ kcal/mol, or when the dihedral pairs option is not used.

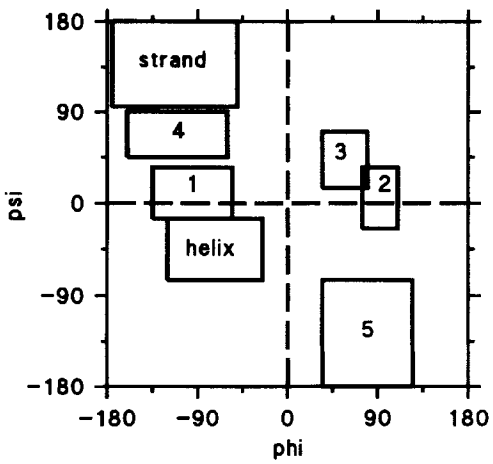

FIGURE 1. Ramachandran ϕ/ψ plot. α-Helical and β-strand rotamer regions are labeled helix and strand, respectively. The five rotamer regions in the β-hairpin class are numbered 1–5.

Results

The calculation of the B5 segment, which is relatively easy to compute, is presented first. The calculation of the H5 segment demonstrates how well program Safd/2 does for a loop with a short end-to-end distance. The best H5 simulation run is characterized in terms of computed acceptance ratio, computed segment hydrogen bonding, and molecular energetics.

CALCULATION OF B5

In the B5 calculation, five equilibration runs, each starting from a different random conformation, are executed. The final conformation of an equilibration run is the starting point for two groups of four production runs in which NCON is 1500 conformations per step. The groups of production runs employ different random number generator seeds. Results of the B5 calculation are given in Table I. Production runs that follow the two best equilibration runs are distinguished by the heading equilibration 3 and equilibration 5 in Table I. Column 2 is the final total energy [Eq. (1)]. The rms backbone deviations are given under column 3. Columns 4 and 5 give the number of backbone–backbone and side chain–side chain H bonds, respectively. All H bonds that involve the computed segment are counted. The H bond criteria are given in footnote d. Initial Boltzmann temperature factors of prediction runs 1–4 under

TABLE I.
Calculation to Compute Segment B5 in BPTI.

Run[a]	E_{total}[b]	rms Backbone Deviation[c]	No. H bonds[d] B/B	S/S
Min	−121.9	0.11	5	1
Equilibration 3				
1a	−118.3	0.72	3	0
2a	−114.1	1.00	3	0
3a	−125.9	0.67	4	0
4a	−116.1	0.75	2	0
1b	−110.3	1.17	2	1
2b	−120.7	0.94	3	0
3b	−116.6	0.74	3	0
4b	−129.7	1.03	5	0
Equilibration 5				
1c	−89.63	1.15	2	1
2c	−113.5	0.79	4	0
3c	−128.3	0.79	4	0
4c	−129.9	0.56	5	1
1d	−110.6	0.90	4^{1x}	1
2d	−130.9	0.79	3	0
3d	−129.9	0.51	4	0
4d	−129.7	0.73	3	0

Analyses of the final conformations in the calculation to compute the β-strand segment in BPTI. Starting from the conformation at the end of the indicated equilibration, groups of multiple prediction plus refinement simulations are computed. NCON is 1500 conformations per step. All rms overlap deviations, eq. (4) are less than 0.1 Å.

[a] Min means X-ray structure refined by energy minimization. See text for SA protocol.

[b] Total energy [see eq. (1)] The final total energies of the LECs are underlined.

[c] Computed segment backbone deviation from the X-ray conformation (N, C^{α}, C^{β}, C', and O; in Ångstroms). Backbone heavy atoms in the protein except for those in the computed segment are superposed on the X-ray structure. In this study, rms deviations between atoms superposed are: backbone, ≤ 0.14 Å, side chain, ≤ 0.37 Å; all, ≤ 0.25 Å.

[d] Number of H bonds in which at least one of the atoms in the H bond belongs to the computed segment. H-bond criteria are: H-bond length is between 1.1 and 2.5 Å, and H-bond angle is in the range 180° ± 45°. Columns labeled B/B and S/S give the numbers of backbone-backbone and side chain-side chain H bonds, respectively. Superscript 1x indicates that one of the H bonds is not present in the refined native conformation.

column 1 are 500, 200, 100, and 75 kcal/mol, respectively. The two groups of production runs that follow an equilibration run are distinguished by the letters "a" and "b" appended to the run numbers. Min under column 1 means refined native X-ray conformation by means of energy minimization.

The final total energies of seven LECs are underlined in Table I. The final total energies of 5 LECs are within 2.6 kcal/mol of the LEC final total energy (−130.9 kcal/mol; run 2d), and one conformation is within 5.0 kcal/mol. The rms backbone deviations of run 4b and run 4c final conformations, which have all five native backbone–backbone H bonds, are 1.03 and 0.56 Å, respectively. The rms backbone deviations of the remaining LECs are between 0.51 and 0.79 Å. A similar B5 calculation computed with NCON equal to 600 produced four LECs with an average rms backbone deviation of 0.59 ± 0.24 Å. As a measure of efficiency, the calculation evaluated 0.40 million trial conformations per LEC.

The B5 calculation was used to evaluate the use of initial equilibration runs. In the test calculation, a group of production simulations starting from a random loop conformation are computed, i.e., equilibration runs are not employed. Initial Boltzmann temperature factors of production runs are 2000, 1000, 500, 200, 100, and 75 kcal/mol. Groups of production runs are computed for NCON equal to 600, 1500, 2100, 3000, 4500, and 6000 conformations per step. The B5 equilibration test calculation produced only 2 LECs for which the average rms backbone deviation is 0.65 ± 0.11 Å. The number of trial conformations evaluated per LEC is 3.33 million. The calculation that included initial equilibration runs is 8.4 times more efficient.

CALCULATION OF H5

The H5 calculation consists of ten equilibration runs, each starting from a different random conformation. The final conformation of an equilibration run is the starting point for two groups of four production runs in which NCON is 3000. The final conformation of one equilibration run led to the LECs. Two additional groups of four production runs starting from the final conformation of the best equilibration run are computed. Results of the H5 calculation are given in Table II. Columns 2, 3, 4, and 6 are the same as columns 2, 3, 4, and 5 in Table I. Column 5 gives the number of backbone–side chain H bonds. The four groups of production runs are distinguished by letters appended to the run numbers. Initial temperature factors of prediction runs 1–4 are 500, 200, 100, and 75 kcal/ mol, respectively.

The LECs in the H5 calculation are obtained in runs 2a and 3b in Table II. The final total energies of runs 2a and 3b are −439.5 and −437.5 kcal/mol,

TABLE II.
Calculation to Compute Segment H5 in BPTI.

Run[a]	E_{total}[b]	rms Backbone Deviation[c]	No. H bonds[d] B/B	B/S	S/S
Min	−368.2	0.09	4	2	1
1a	−387.6	2.49	1	1*	3*
2a	−439.5	0.58	3	2*	4*
3a	−394.9	2.01	1	0	3*
4a	−417.6	1.68	2	3*	4*
2b	−408.0	1.82	2	2^{1x}	4*
3b	−437.5	0.27	4	2^{1x}	3*
4b	−395.7	2.33	1	1*	2*
1c	−424.8	1.26	3^{1x}	3*	4*
3c	−420.7	3.46	1	1*	2^{1x}
4c	−416.3	0.90	2	1*	4*
2d	−386.3	3.73	1	1*	3^{2x}
3d	−393.7	2.63	1	0	1
4d	−399.9	1.89	2	3^{2x}	2*

Starting from the conformation at the end of the equilibrium simulation, groups of multiple prediction plus refinement simulations are computed. NCON is 3000 conformations per step. The final conformation of each refinement simulation is analyzed. Simulations with final rms overlap deviations less than 0.1 Å are analyzed.

[a] Min means X-ray structure refined by energy minimization. See text for SA protocol.

[b,c] See footnotes b and c of Table I.

[d] See footnote d of Table I. Columns labeled B / B, B / S, and S / S give the numbers of backbone–backbone, backbone–side chain, and side chain–side chain H bonds, respectively. Superscripts 1x, 2x, and * indicate that one, two, and all computed H bonds, respectively, are not present in the refined native conformation.

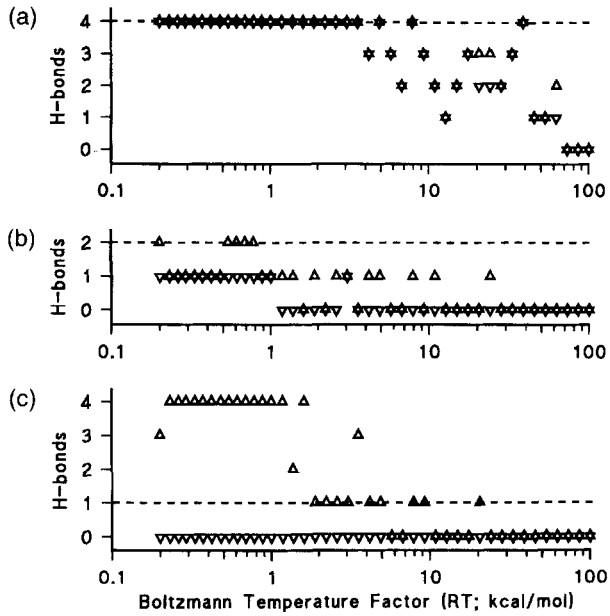

FIGURE 2. Number of H bonds in the helical segment, H5, is plotted against Boltzmann temperature factor for run 3b in Table II. H bonds plotted are (a) backbone–backbone, (b) backbone–side chain, and (c) side chain–side chain. (*) all the computed H bonds are correct. (△) Number of computed H bonds. (▽) Number of computed H bonds that are correct. (---) Number of H bonds in the refined native conformation.

respectively. The rms backbone deviations of runs 2a and 3b are 0.58 and 0.27 Å, respectively. The final total energy of the next lowest LEC is 14.7 kcal/mol higher than the LEC. All the native backbone–backbone H bonds are present in the run 3b final conformation.

CHARACTERIZATION OF BEST H5 SIMULATED ANNEALING RUN

H bonding, molecular energetics, and acceptance ratios during the best H5 simulated annealing run, run 3b in Table II, are characterized here. Characteristics of simulation runs 2a and 3b are for all practical purposes identical.

Computed segment H bonds at the end of each MCSA step are analyzed in Figure 2. The number of backbone–backbone H bonds, backbone–side chain H bonds, and side chain–side chain H bonds are plotted against *RT* in Figure 2a, b, and c, respectively. The simulation proceeds from right to left, from large *RT* to small *RT*. H-bond criteria are given in footnote d to Table I. The number of computed segment H bonds are represented by triangles with the apex pointing up. A triangle with the apex pointing down represents the number of correct computed segment H bonds–H bonds present in the refined native conformation. A star in Figure 2 indicates that all observed H bonds are correct. A horizontal dashed line is the number of H bonds in the refined native structure. Figure 2 suggests that the segment main chain is fixed in the native fold before side chain H bonds are stable. According to Figure 2a, backbone–backbone H bonds stabilize in the 70–4 kcal/mol *RT* range. According to Figure 2b and c, H bonds that involve side chains stabilize in the 24–1 kcal/mol range.

Molecular energetics during the best H5 simulated annealing run are plotted in Figure 3. The solid line, broken line, and dashed line connect the computed total energy, local region energy, and weighted polypeptide chain continuity constraint of the conformation at the end of each simulated

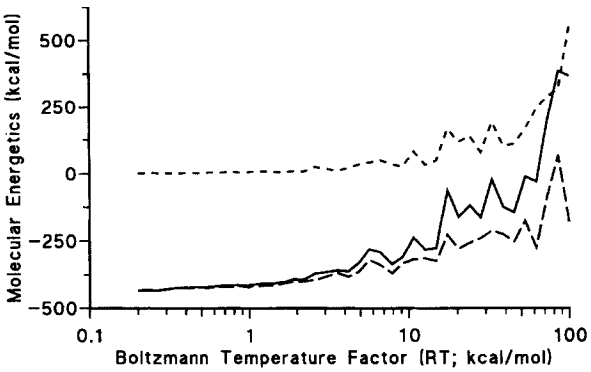

FIGURE 3. Molecular energetics of H5 are plotted against Boltzmann temperature factor for run 3b of Table II. Total energy, local region energy, and weighted polypeptide chain continuity constraint at the end of each simulated annealing step are connected by solid line, broken line, and dashed line, respectively.

annealing step. The total energy is the sum of the local region energy and weighted chain continuity constraint. Total energy fluctuations of 100 kcal/mol occur in the RT range that main chain H bonds stabilize. Below $RT = 4$ kcal/mol, main chain H bonds are intact; H bonds that involve side chains stabilize; and total energy fluctuations are about 10 kcal/mol.

Figure 4a and b are plots of computed acceptance ratios based on random walk and rotamer trial conformations, respectively. Computed acceptance ratios based on the number of accepted main chain and side chain trial conformations are given by open triangles and open circles, respectively. Euler angle and translation acceptance ratios are plotted in Figures 4c and d, respectively. Horizontal dotted lines in Figure 4 are acceptance ratio limits. The computed main chain (Fig. 4a), Euler angle, and translation acceptance ratios drop below the 25% limit several times during the simulation. Each time the acceptance ratio drops below the limit, the corresponding perturbation is lowered; and the acceptance ratio computed at the end of the next step is larger. The acceptance ratio based on on *ω-trans* rotamer trial conformations is plotted between $RT = 50$ and 1 kcal/mol in Figure 4b. The initially large side chain acceptance ratios in Figures 4a and b gradually decrease during the simulation. The sharp drop in the side chain acceptance ratio at $RT = 1.0$ kcal/mol in Figure 4a is due to the large side chain perturbation employed in the refinement run. In Figure 4b, the side chain rotamer acceptance ratio increases

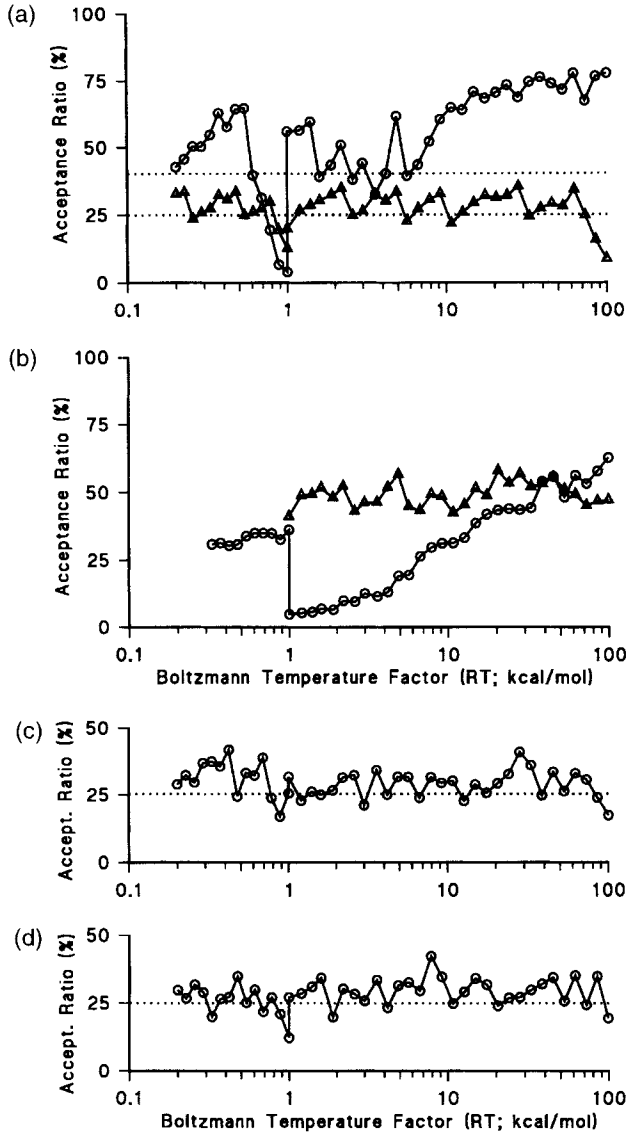

FIGURE 4. Acceptance ratios in the H5 simulation are plotted against Boltzmann temperature factor for run 3b in Table II. Types of coordinates explored are (a) main chain (△) and side chain (○), (b) main chain rotamer (ϕ/ψ and ω; △), and side chain rotamer (○), (c) Euler angle, and (d) translation. (···) Acceptance ratio limits.

sharply at $RT = 1.0$ kcal/mol where side chain dihedral angles with one rotamer range are varied.

PROGRAM ENHANCEMENTS TEST CALCULATIONS

Program enhancements evaluated in test calculations are: altering the rate in which main chain versus side chain dihedral angles are selected to

generate trial conformations; varying ϕ/ψ main chain and χ_1/χ_2 side chain dihedral angles in pairs; the use of the main chain rotamer library; the use of the side chain rotamer library; and changing the location of the segment origin during the simulated annealing. A test calculation consists of six production runs starting from random initial conformations of the B5 segment. In a B5 test calculation, one of the program options is turned off and the calculation is repeated. The benchmark B5 calculation employs all program enhancements. Based on rms overlap deviations, eq. (4), more than four out of six final conformations are broken in each test calculation. Computed lowest final total energies are between 8.3 and 28.9 kcal/mol higher than the benchmark LEC total energy. B5 test calculations did as well as the benchmark, in terms of rms backbone deviation.

Discussion

The difficult part of the β-strand folding simulation is overcome in the equilibration run. In the final conformations of the two best B5 equilibration runs, the number of ϕ/ψ dihedral angle pairs that lie in the β-strand region of the Ramachandran plot (Fig. 1) are four out of five and five out of five. None of the ϕ/ψ pairs in the conformation at the end of the best H5 equilibration run lie in the helix region. The fact that the H5 conformational search did not easily distinguish between native and nonnative conformations is likely due to the short end-to-end distance of the α-helix N-terminus segment. Current studies in the first author's laboratory focus on the conformational search problem at Boltzmann temperature factors in which the main chain adopts the native fold.

Based on the measured hydrogen exchange (HX) rate constants (corrected for amino acid type) in the solution structure of BPTI,[42] no fraying at the helix N-terminus is experimentally observed. All four H-bonded peptide NH in the H5 segment are equally protected from HX. Results of the H5 calculation are consistent with the HX rate data. For simulated annealing temperature factors as large as 4 kcal/mol in the best H5 simulation run, all four native main chain–main chain H bonds are intact.

Calculations in this study are more reliable, reproducible, and efficient compared to corresponding calculations in the previous study.[20] The calculation is reliable when the rms backbone deviation of the LECs from the native X-ray conformation are 1.0 Å or less. The calculation is reproducible if several computed LECs are close to the native structure. The calculation efficiency is the number of trial conformations evaluated per LEC.

In the B5 calculation, 11 LECs are computed. The average rms backbone deviation is 0.68 $\pm$ 0.20 Å. The number of trial conformations evaluated per LEC is 0.40 million. The B5 calculation of the previous study computed only 1 LEC for which the rms backbone deviation is 1.11 Å. And, 9.72 million trial conformations are evaluated per LEC. Compared to the previous study, the B5 calculation here is 24.4 times more efficient.

In the H5 calculation, 2 LECs are computed and the average rms backbone deviation is 0.43 $\pm$ 0.15 Å. The computational efficiency is 6.44 million trial conformations per LEC. The calculation here is a factor of 2.8 times faster than the H5 calculation of the previous study.

In the MCSA calculation by Collura et al.,[19] the average rms backbone deviation of the computed seven-residue N-terminus helical segment in BPTI, which overlaps with H5 in this study, is 0.5 Å. The calculation efficiency is 1.05 million trial conformations per LEC compared to 6.44 million trial conformations per LEC in our study. The method employed by Collura and colleagues utilizes a matrix of energy second derivatives to guide the conformational search. The calculation of energy derivatives is not included in the calculation of computational efficiency here. MCSA loop calculations in our previous study are compared to loop calculations by Bruccolieri and Karplus[21] and by Fine and collaborators.[22,23]

Conclusions

MCSA program enhancements and the use of initial equilibration runs significantly improve the calculation of surface loops in proteins starting from random conformations. For a five residue helical loop, which has a short end-to-end distance, the computational effort is reduce by a factor of 3. For a five-residue β-strand loop, which has a large end-to-end distance, the computational effort is reduce by a factor of 24. Average rms backbone deviations of the computed five-residue segments from the native are less than 0.7 Å. All the native backbone–backbone H bonds are intact in the best computed conformations.

Acknowledgments

This research was supported in part by an NIH research grant and a grant from the Pittsburgh Supercomputing Center through the NIH Division of Research Resources Cooperative Agreement 1P41 RR06009-01 and through a grant from the National Science Foundation Cooperative Agreement ASC-8500650. We thank Professor Harvey Rubin for the use of his computer workstation. We appreciate helpful comments by Dr. Kuo-Chen Chou and by referees.

References

1. P. Argos, M. G. Rossmann, and J. E. Johnson, *Biochem. Biophys. Res. Commun.*, **75**, 83 (1977).
2. P. C. Weber and F. R. Salemme, *Nature*, **287**, 82 (1980).
3. J. S. Richardson, *Adv. Protein Chem.*, **34**, 167 (1981).
4. S. T. Rao and M. G. Rossmann, *J. Mol. Biol.*, **76**, 89 (1973).
5. M. G. Rossmann, D. Moras, and K. W. Olsen, *Nature (London)*, **250**, 194 (1974).
6. L. M. Amzel and R. J. Poljak, *Annul Rev. Biochem.*, **48**, 961 (1979).
7. D. W. Banner, A. C. Bloomer, G. A. Petsko, D. C. Philips, and I. A. Wilson, *Biochem. Biophys. Res. Commun.*, **72**, 146 (1976).
8. A. M. Lesk, C. I. Branden, and C. Chothia, *Proteins*, **5**, 139 (1989).
9. G. K. Farber and G. A. Petsko, *Trends Biochem. Sci.*, **15**, 228 (1990).
10. J. S. Richardson, K. A. Thomas, B. H. Rubin, and D. C. Richardson, *Proc. Natl. Acad. Sci. USA*, **72**, 1349 (1975).
11. R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing, *J. Mol. Biol.*, **213**, 899 (1990).
12. S. W. Englander and N. R. Kallenbach, *Q. Rev. Biophys.*, **16**, 521 (1984).
13. S. W. Englander, J. J. Englander, R. E. McKinnie, G. K. Ackers, G. J. Turner, J. A. Westrick, and S. J. Gill, *Science*, **256**, 1684 (1992).
14. K. C. Chou and L. Carlacci, *Protein Eng.*, **4**, 661 (1990).
15. S. R. Wilson and W. Cui, *Biopolymers*, **29**, 225 (1990).
16. C. Lee and S. Subbiah, *J. Mol. Biol.*, **217**, 373 (1991).
17. C. Lee and M. Levitt, *Nature*, **352**, 448 (1991).
18. J. Higo, V. Collura, and J. Gamier, *Biopolymers*, **32**, 33 (1992).
19. V. Collura, J. Higo, and J. Garnier, *J. Protein Sci.*, **2**, 1502 (1993).
20. L. Carlacci and S. W. Englander, *Biopolymers*, **33**, 1271 (1993).
21. R. E. Bruccoleri and M. Karplus, *Biopolymers*, **26**, 137 (1987).
22. R. M. Fine, H. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal, *Proteins*, **1**, 342 (1986).
23. P. S. Shenkin, D. L Yarmush, R. M. Fine, H. Wang, and C. Levinthal, *Biopolymers*, **26**, 2053 (1987).
24. K. C. Chou, G. Némethy, M. Pottle, and H. A. Scheraga, *J. Mol. Biol.*, **205**, 241 (1989).
25. L. Carlacci and K.-C. Chou, *Biopolymers*, **30**, 135 (1990).
26. L. Carlacci and K.-C. Chou, *Protein Eng.*, **3**, 509 (1990).
27. L. Carlacci and K.-C. Chou, *Protein Eng.*, **4**, 225 (1990).
28. K. C. Chou and L. Carlacci, *Proteins*, **9**, 280 (1991).
29. L. Carlacci, K.-C. Chou, and G. M. Maggiora, *Biochemistry*, **30**, 4389 (1991).
30. J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, **193**, 775 (1987).
31. J. March, In *Advanced Organic Chemistry*, R. H. Summersgill and A. T. Vinnicombe, Eds., McGraw-Hill, New York, 1977, pp. 71, 113.
32. B. L. Sibanda, T. L. Blundell, and J. M. Thornton, *J. Mol. Biol.*, **206**, 759 (1989).
33. (a) F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535, (1977); (b) E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, *Protein Data Bank in Crystallographic Databases —Information Content, Software Systems, Scientific Applications*, F. H. Allen, G. Bergerhoff, and R. Sievers, Eds., Data Commission of the International Union of Crystallography, Bonn/ Cambridge/ Chester, 1987, p. 107.
34. IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, **9**, 3471 (1970).
35. F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga, *J. Phys. Chem.*, **79**, 2361 (1979).
36. G. Némethy, M. S. Pottle, and H. A. Scheraga, *J. Phys. Chem.*, **87**, 1883 (1983).
37. Molecular Simulations, Inc., 200 Fifth Avenue, Waltham, MA 02154. CHARMm is a trademark of Molecular Simulations.
38. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. J. Teller, *Chem. Phys.*, **21**, 1087 (1953).
39. S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, *Science*, **220**, 671 (1983).
40. L. T. Whille, *Nature*, **324**, 46 (1986).
41. C. Wilson and S. A. Doniach, *Proteins*, **6**, 193 (1989).
42. G. Wagner and K Wüthrich, *J. Mol. Biol.*, **160**, 343 (1982).